Learnable Diffusion Framework for Mouse V1 Neural Decoding

Kaiwen Deng[1, *, +], Peter S. Schwendeman[1, 2, *], Yuanfang Guan[1, 3, +]

1. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, 48105
2. Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48105
3. Department of Internal Medicine, University of Michigan, Ann Arbor, MI, 48105

* These authors contributed equally
+ Contact: dengkw@umich.edu, gyuanfan@umich.edu

Abstract

Decoding visual stimuli from neural signals is an essential step toward understanding how sensory information is represented in the brain. While most existing approaches reconstruct visual stimuli from human functional magnetic resonance imaging (fMRI), utilizing calcium imaging in mice opens the door to single-neuron-level insights into non-primate visual systems with distinct spectral sensitivities. Here, we present Sensorium-Viz, a diffusion-based framework specifically designed for decoding activity in the mouse primary visual cortex. The model is the first to reliably reconstruct complex, high-resolution images from previously unseen single-neuron responses. At its core, Sensorium-Viz integrates advanced training strategies and architectural innovations for neuron-to-image decoding. Incorporating synthetic responses during training boosts reconstruction quality by over 30% and enables efficient cross-mouse generalization through fine-tuning. Combining a Diffusion Transformer (DiT) with a spatial embedder tailored for neuronal responses further allows Sensorium-Viz to surpass leading fMRI-based methods by up to 13.16% on both pixel- and content-level metrics. Analysis of the neural responses and corresponding reconstructions reveals that neurons sensitive to low-level visual features form the primary basis of V1's representation of external stimuli. These findings establish Sensorium-Viz as a biologically grounded and technically robust tool for vision decoding.

Introduction

Reconstructing visual stimuli from brain activity has long been a focus of neuroscience and machine learning research, offering valuable insights into how sensory information is processed in the brain and paving the way toward more advanced brain-computer interfaces [1–3]. Recent advancements in neuroimaging and deep learning have enabled the more accurate mapping of brain signals to visual stimuli, thereby expanding the potential for reconstructing complex, high-quality images from neural data.

Most prior research on visual reconstruction has utilized functional magnetic resonance imaging (fMRI) data from human brains to regenerate colored visual stimuli. fMRI enables non-invasive observation of population-level brain activity by tracking blood oxygen levels, revealing correlations between neural responses and visual information [4–6]. These approaches have primarily focused on retrieving latent representations from fMRI signals and using them together with generative adversarial networks (GANs) to guide image reconstruction [7–9]. However, the most recent developments in generative

AI have further advanced this field, with diffusion models standing out for their exceptional ability to generate semantically rich reconstructions from human brain data [10,11].

Despite the success of fMRI-based reconstruction in humans, much less is known about how visual experiences can be reconstructed in other species and with alternative recording modalities. Studying such systems is crucial for revealing both general principles and species-specific features of visual processing. In particular, mice have long served as model organisms for non-primate visual systems [12,13], and advances in two-photon calcium imaging with craniotomy now allow researchers to record activity from individual neurons, providing a level of detail not accessible with population-level measures [14]. Leveraging these capabilities, a pioneering study by Yoshida and Ohki (2020) demonstrated that images could be reconstructed from mouse calcium signals using a cell-selection linear model on Gabor filter–based features [15]. While this investigation demonstrated the feasibility of neuron-to-image decoding, it relied on contrast-enhanced stimuli that were downsampled to 32x32 pixels for reconstruction, along with handcrafted Gabor kernels, which are known to capture mostly low-level edges and textures [16]. Other studies explored non-linear models and convolutional neural networks (CNNs), but these efforts were similarly constrained by low-resolution stimuli and achieved only moderate performances [17,18].

At the same time, attempts to directly train state-of-the-art human fMRI reconstruction models, such as MinD-Vis, on mice V1 single-neuron responses similarly fail to produce semantically meaningful images or accurate pixel-level alignment [10] (Supplementary Figures 1A, 1B). Together, these limitations emphasize the need for models specifically designed for non-human visual systems and raise several compelling questions: How can advanced AI architectures be adapted to calcium–imaging–based decoding for more complex visual stimuli? How can the limited data available per subject be leveraged efficiently to meet the large-sample requirements of AI models? And how do the behaviors of such models align with the biological understanding of visual coding in V1?

Motivated by these questions, we introduce Sensorium-Viz, an image reconstruction framework tailored for calcium imaging data from the mouse primary visual cortex. Sensorium-Viz combines a lightweight spatial embedding module for neuron responses with a Diffusion Transformer (DiT) model, which generates images conditioned on these embeddings. We demonstrate two key advances. First, we describe a synthetic response augmentation strategy that improves reconstruction quality and enables efficient cross-subject generalization. Second, we show that our architecture outperforms state-of-the-art fMRI-based approaches under equivalent training and evaluation conditions. Beyond performance, we demonstrate that specific feature-selective neurons, rather than global population activity, play a central role in V1's representation of visual stimuli. Together, these results establish Sensorium-Viz as a robust tool for high-fidelity visual decoding in animal models, offering new opportunities for probing the neural basis of complex visual experiences.

Results

Sensorium-Viz reconstructs stimuli via DiT with spatially embedded neuron responses

Here, we describe a novel reconstruction model for calcium image data, which we later use to investigate how signals from the primary visual cortex in mouse brains represent and process visual stimuli. To train the model, we used training data obtained from the Sensorium Challenge, which was

designed for neural encoding tasks (*i.e.,* predicting the corresponding neural signals from given visual stimuli). This dataset consists of neuronal recordings from five awake, head-fixed mice presented with grayscale stimuli derived from ImageNet images[19]. These neuronal responses were recorded from layers 2 and 3 (L2/3) of the right primary visual cortex (V1). The spiking activity of neurons was captured via two-photon calcium imaging and quantified as relative fluorescence changes. Neuronal responses were accumulated between 50 and 550 ms after stimulus onset using a boxcar window. Each neuron's response to a given stimulus was represented by a single value. The number of recorded neurons varied across mice (Table 1). Additionally, the anatomical coordinates of each neuron relative to the pial surface were also recorded (Figure 1A).

Table 1. Experimental information of the mouse in the Sensorium Challenge datasets.

Mouse ID	Neuron #	Total trials #	Test trials #	Test instance #
21067	8372	5994	998	100
22846	7344	5997	999	100
23343	7334	5951	989	100
23656	8107	5966	993	100
23964	8098	5983	994	100

Sensorium-Viz employs a Diffusion Transformer (DiT) to model the relationship between visual inputs and neural responses (Figure 1D). DiT utilizes an enhanced denoising function through transformer-based blocks, similar to those seen in Vision Transformers (ViTs), and encodes external information as conditions using Adaptive Layer Norm (adaLN) Zero blocks (Figure 1D, Supplementary Figure 2B) [20]. This improved architecture yielded better image generation quality for grayscale images in our preliminary experiments (Supplementary Figures 1C and 1D). Models were trained separately for each mouse on its trials, excluding the test ones (Table 1). As with other diffusion models, training occurred in two stages: forward diffusion, where input stimuli are projected into a low-dimensional latent space and progressively perturbed with Gaussian noise; and reverse denoising, where the denoising function is trained to iteratively remove noise and reconstruct the original input, guided by the neural responses.

Sensorium-Viz receives the image and neuron response data during training and only requires the latter for back-inferencing the input images. Images are resized and center-cropped to a shape of 256×256 to fit the model's requirements (Figure 1C). Center cropping minimizes information loss by preserving the image regions most relevant to neural activity [21]. To encode the neuron responses, we account for the noisy nature of neuronal activity and its spatial dependence within the brain. In our previous study, we found that grid-based embedding methods have shown strong ability in summarizing neuron signals while preserving their spatial properties[24]. Thus, we utilized a similar strategy that embedded the responses into 32×32 grids using an Inverse Distance Weighting (IDW) algorithm, which uses interpolated 2D neuron positions as XY coordinates, normalized to the range of -1 to 1. The resulting spatially structured

response maps are further processed by a grid response projector network, which encodes them as conditions to guide the DiT model's reconstructions (Figure 1C, Supplementary Figure 2A).

To validate the embeddings, we compared signal correlations across repeated presentations of each test stimulus in individual mice. Raw neural responses showed an average correlation of 0.2304 (±0.09, average p-value = 0.0246) between repetitions of the same stimuli and 0.0149 (±0.02) between different test stimuli (*i.e.,* test images with different content). In contrast, the embedded responses exhibited within-stimulus correlations of 0.2551 (±0.10, avg. p-value = 0.0571) and between-stimulus correlations of 0.0224 (±0.03). These results suggest that the embeddings preserve critical information from the raw signals, necessary to distinguish between different visual stimuli (Figure 1E, Supplementary Figures 3 and 4).

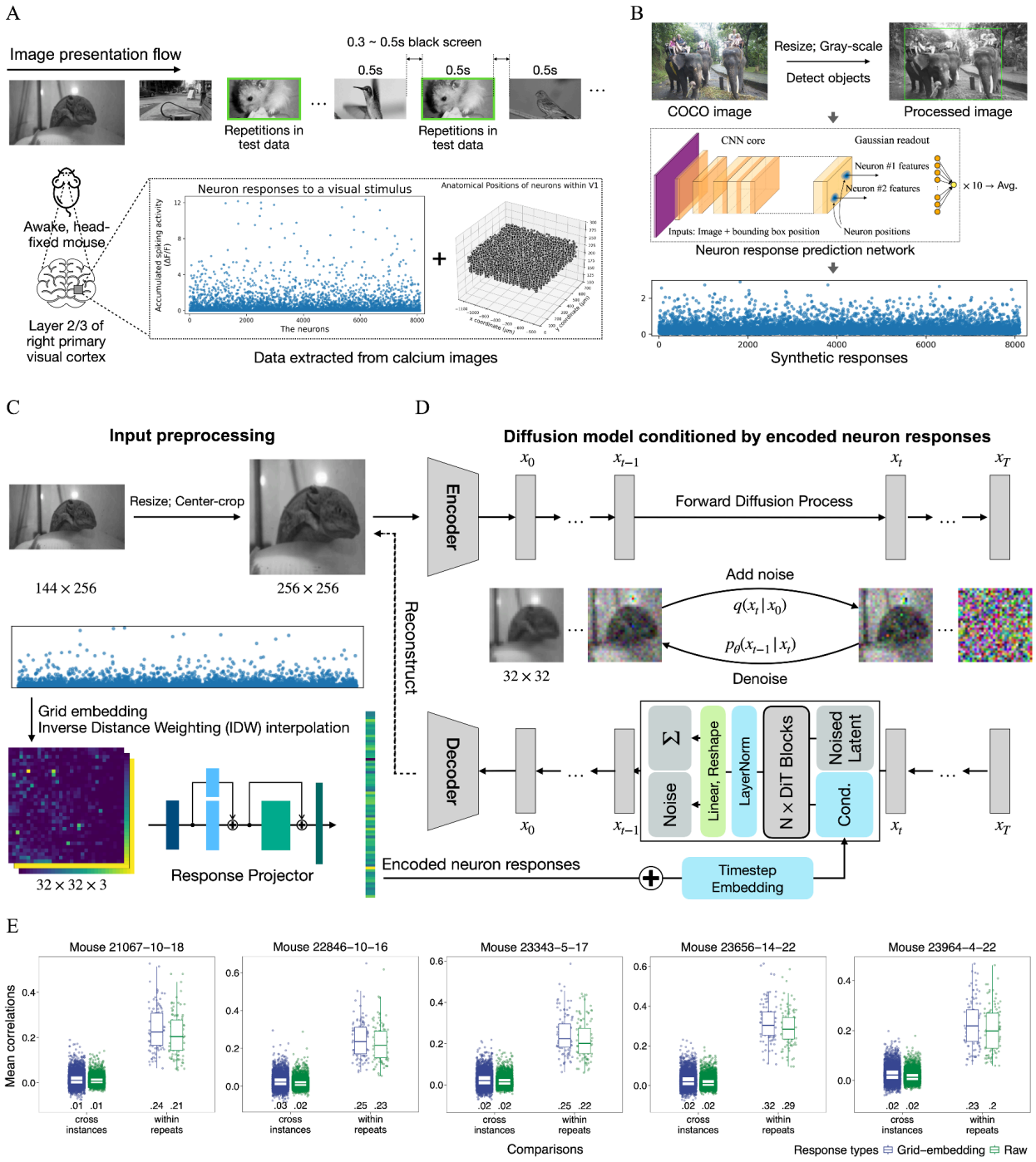

Figure 1. Utilize the Diffusion Transformer (DiT) to reconstruct the visual stimuli conditioned by the neuron responses. (A) describes how the images were presented to the mice in the datasets and the data types used to train the models. The images were shown to awake, head-fixed mice in a flow with 0.5-second presentation times, followed by 0.3-0.5 seconds of black screens. There were 100 instances used as the test data in the original datasets, which were repeated 10 times and mixed with the other images in the flow (the image with the green border in this figure is an example). The neuron responses to a specific image, along with their anatomical positions, were extracted from V1 of the mouse brains. (B) shows the pipeline for generating synthetic responses from COCO image sets. The processed images and the object positions in the images, as well as the neuron coordinates of the mice,

were given to our pre-trained neuron signal prediction models to generate the synthetic responses specifically for images and mice. There were 10 pre-trained models, and the final output was their average. (C, D) draw the complete pipeline of our method. (C) indicates our preprocessing on the image and neuron responses, including resizing the input images and embedding and encoding the neuron response to fit the input shapes of DiT. (D) briefly describes the overall structure of the diffusion model. (E) presents the results of signal correlations comparing the repetitions within the same instances and across different instances. The blue points and boxes are for the responses after spatial embeddings. The green ones are the raw responses. Each point represents the average correlation among the repetitions.

Synthetic responses and stimulus repetition improve the reconstruction qualities

We first assessed its reconstruction qualities using test data, which included grayscale images with mostly 10 repetitions of 100 unique instances (Table 1). Five models were trained separately for each of the five mice and evaluated using two metrics (Figure 2A): Pearson's correlation among pixels (pixel-wise correlation) and the structural similarity index (SSIM), as well as their pairwise similarities, following the reports in previous works (Figure 2B) [8,22]. These metrics quantify the qualities at the pixel level and of the overall image content. Pixel-level mean squared error (MSE) and Learned Perceptual Image Patch Similarity (LPIPS) were also reported as additional evaluation metrics [23]. Additionally, since the diffusion model generates images from random noise, we performed four reconstructions per target image. We reported the standard deviations of these metrics to evaluate the stability of the inference.

The task of reconstructing images from neural activity poses significant challenges due to two core limitations of the data. First, neuronal responses are highly noisy and sparsely activated: only a small subset of neurons responds reliably to a given stimulus, while the rest exhibit seemingly random fluctuations. Second, our dataset is relatively small, making it hard for our models to learn generalizable patterns. To mitigate these issues, we apply two key strategies during training and inference to enhance the reconstructions. First, during inference, we reduce response variability by averaging across the 10 repetitions for each image. Given the low response correlation within instances (Figure 1E), this approach stabilizes the representation of each stimulus. Second, inspired by the training strategy of DALL-E 3[24], we augment our dataset with additional training images paired with synthetic responses. To perform this augmentation, we collected 82,784 additional images from the COCO dataset and generated corresponding synthetic neural responses using our previously trained neuron signal prediction model (Figure 1B) [25]. It consisted of a shared module for extracting features from images and the separated Gaussian readout modules that mapped the image features to neuronal responses for each mouse. The training datasets also consist of the trials in Table 1, excluding the test trials.

Applying these strategies resulted in significant improvements in reconstruction performance. After merging the repetitions, the median pixel-wise correlation and SSIM scores increased by 52.05% and 53.52% on average, respectively. Pairwise similarities also improved by 15.91% and 11.22% (Figure 2C). MSE and LPIPS, along with their pairwise similarities, also showed significant improvements (Supplementary Figure 6). Incorporating 80,000 additional synthetic samples further enhanced reconstruction quality, leading to 30.43% and 38.20% higher median pixel-wise correlation and SSIM scores, as well as 17.91% and 12.96% improvements in pairwise similarities (Figure 2C). Visual inspection of the reconstructions confirmed these quantitative results. Models trained with more synthetic samples produced images that more closely resembled the ground truth (Figure 2D). Similar improvements were also observed for MSE and LPIPS (Supplementary Figure 6).

Among the five mice, the final metric scores of the best-performing model were 0.4581 for pixel-wise correlation (min: 0.4367 - max: 0.5092, average std. among the four reconstructions for one image: 0.0069), 0.4235 for SSIM (0.4034 - 0.4722, std.: 0.0060), 0.9191 for pairwise similarity in pixel-wise correlation (0.8897 - 0.9339, std.: 0.0059), and 0.7309 for pairwise similarity in SSIM (0.7154 - 0.7532, std.: 0.0058) (Figure 2B, Supplementary Figures 9 - 13).

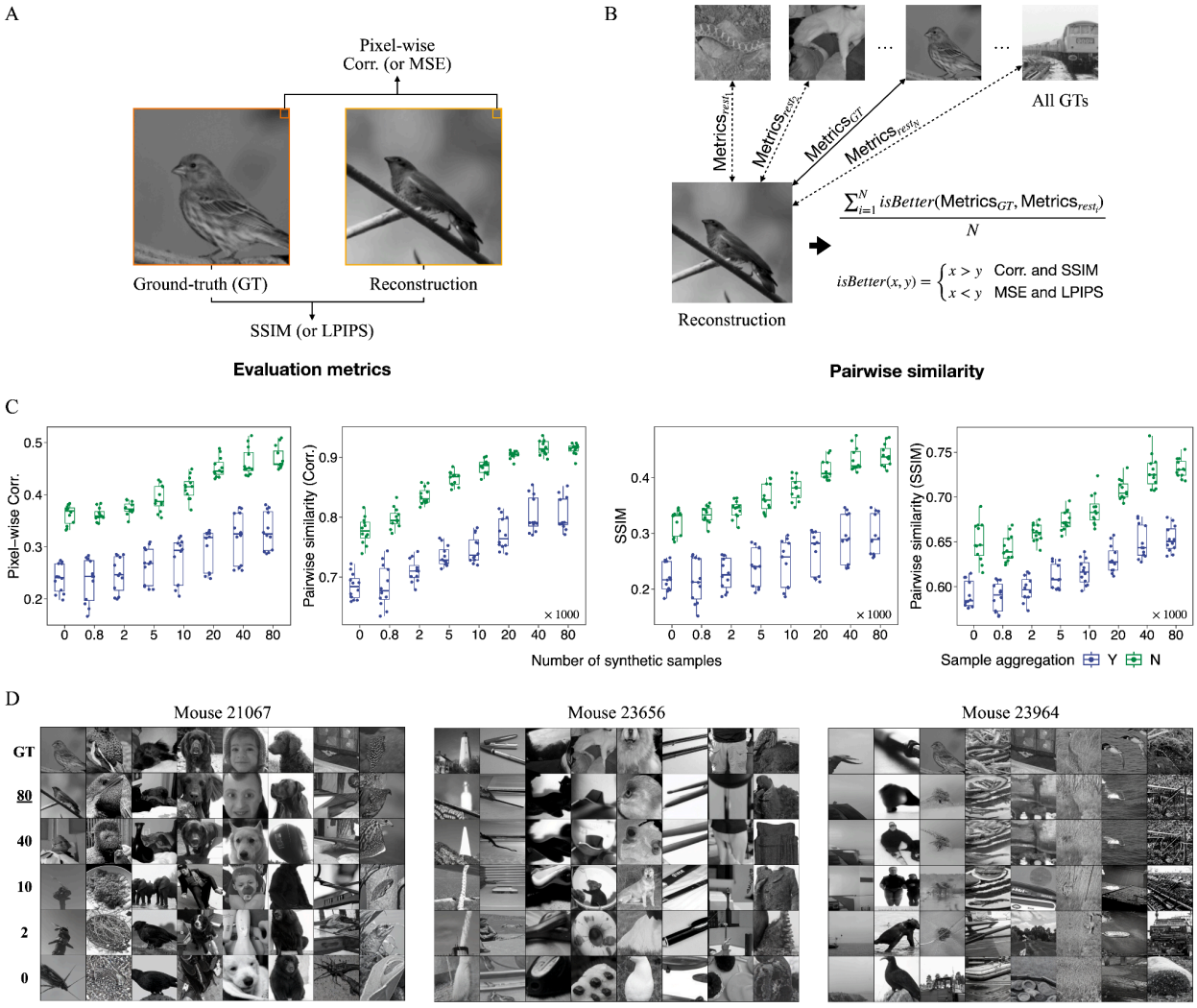

Figure 2. Evaluations of the reconstruction qualities. (A) Overview of the evaluation metrics. SSIM and LPIPS are calculated for the entire image, while MSE and correlations are computed pixel by pixel. (B) Illustration of the pipeline used to compute pairwise similarity for a given metric. The process involves generating scores between the reconstruction and all other ground truths, expecting the metric score for the true ground truth to be the highest. (C) Correlation scores ("Corr.") and SSIM plots, along with their pairwise similarities, for models trained with different amounts of synthetic data. The images reconstructed from aggregated ("Y") or separated ("N") neuronal responses are compared. The numbers on the x-axis represent the real synthetic sample numbers divided by 1,000. (D) Sample reconstructions randomly selected from the high-LPIPS-score results from three mice, compared with their corresponding ground truths (first row). Reconstructions are shown for various numbers of synthetic samples (listed

on the left, consistent with the x-axis labels in panel C). All reconstructed images are from the first generation. Full reconstruction results of these three mice can be found in Supplementary Figures 9, 11, and 13.

Sensorium-Viz outperforms the previous SOTA fMRI method

Due to the limited availability of open-source approaches for calcium imaging-based reconstructions in mice, we evaluated our model by adapting MinD-Vis—a state-of-the-art model originally designed for human fMRI data—comparing its performance on data from three mice in our dataset [10]. MinD-Vis consists of an fMRI embedding module (SC-MDM) and a conditioned latent diffusion module (DC-LDM). Considering the differences in data types, we first adjusted the settings of the SC-MDM module, including the generation of non-negative outputs and the incorporation of neuron position information. We embedded the neural responses of the mice using the settings that achieved their best performance, with an average correlation between recovered and original signals of 0.7228 (Supplementary Figure 5). Then, we generated images using the same training and inference strategies as our method, including both synthetic responses and repetition merging.

As a result, Sensorium-Viz significantly outperformed MinD-Vis, requiring 11.36x less training time and achieving an average 6.73% higher performance across all major metrics (except MSE) and pairwise similarities. Median correlation and SSIM scores for Sensorium-Viz were 0.4584 and 0.4370, respectively, compared to 0.4140 and 0.3862 for MinD-Vis (paired Wilcoxon test p-values < 5e-4). Similarly, median pairwise similarities were 0.9210 (correlation) and 0.7378 (SSIM), while MinD-Vis achieved 0.9013 and 0.7079, respectively (paired Wilcoxon test p-values < 5e-3; Figure 3A, Supplementary Figure 7). Furthermore, our model produced reconstructions with higher visual fidelity, better preserving fine image details (Figure 3B, Supplementary Figures 9-11). These results demonstrate that Sensorium-Viz more effectively captures the distinctive features of calcium imaging data from the mouse visual cortex compared to the fMRI-based model, underscoring the viability of our model as a tool for investigating cortical activity in future studies.

Fine-tuning with synthetic data enables efficient cross-mouse inferences

Due to individual differences among subjects, directly applying a model trained on one mouse to another without adaptation may lead to suboptimal reconstructions. However, training a subject-specific decoder often demands substantial additional neuronal recordings for each new subject, which is both labor-intensive and raises ethical considerations. To explore whether synthetic responses could help alleviate this burden, we investigated a transfer strategy in which a model trained on one mouse's data (the "source") was adapted to another (the "target") using only synthetic responses. These synthetic responses were generated from encoding models previously developed for the target mouse, which themselves had been trained on a limited amount of real data from that subject.

In this setup, we first trained a base model using both real and synthetic data from a given source mouse. We then fine-tuned the base model using only synthetic responses generated for a different target mouse. Finally, the fine-tuned model was used to reconstruct test images for the target mouse.

The results show that cross-mouse fine-tuning yields reconstruction performance approaching that of models trained entirely on the target mouse's data (Figure 3C, Supplementary Figure 8). Both correlation

and SSIM remain relatively high when transferred across mice, with median scores of 0.4037 and 0.3873, respectively. The median pairwise similarities were 0.9018 and 0.7263, respectively. These metrics are comparable to the results from MinD-Vis when trained on the same individual mice, demonstrating that our framework effectively leverages synthetic responses to adapt the learned representation. Visual inspection of the reconstructed images further confirmed these findings. Cross-mouse outputs showed similar pixel distribution patterns to both within-mouse reconstructions and the original ground-truth stimuli, indicating that key visual features were preserved despite differences in the underlying neuronal recordings (Figure 3D). Taken together, these results suggest that synthetic responses can support cross-mouse adaptation by reducing the reliance on extensive new recordings from each subject.

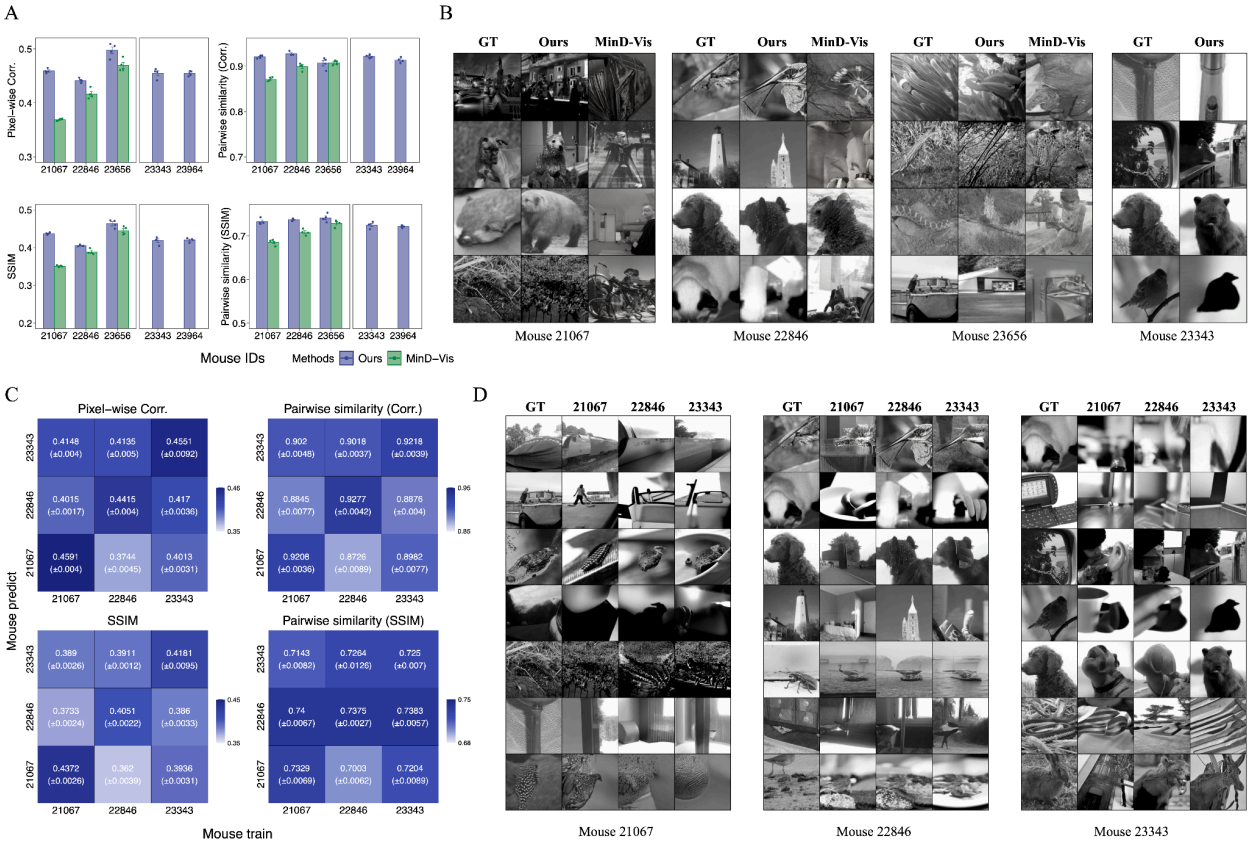

Figure 3. Evaluations of the reconstructions from different methods and cross-mice inferences. (A) Correlation and SSIM scores, along with their pairwise similarities, are used to compare the performance of our method and MinD-Vis. Bars represent the median scores from four generations, while scatter points indicate individual scores for each generation. (B) Visual comparison of ground truths and images reconstructed using our method and MinD-Vis. Our method demonstrates superior reconstruction accuracy at both pixel and semantic levels. (C) Reconstruction performance of transfer-learning models. The x-axis ("Mouse train") specifies the mouse datasets (both real and synthetic) used for training the base models, while the y-axis ("Mouse predict") indicates the synthetic data used for fine-tuning for each mouse. (D) Decoded images comparing models trained on the same mouse and fine-tuned on data from other mice. Titles above the images indicate the mouse providing the base model, while titles below indicate the mouse used for fine-tuning and prediction.

Individual neurons actively responding to image properties significantly contribute to the reconstructions

To better understand the contributions of neuronal activity patterns to decoding, we further analyzed neuron signals at both the individual and population levels, investigating their relationships with image properties and reconstruction metrics. Image properties were quantified by the average spatial information for edge magnitudes, as well as the mean and standard deviation of pixel intensities for brightness and contrast, respectively. These analyses evaluated whether the model's reliance on neuronal activity aligns with known response patterns in V1.

Image properties were quantified by average edge magnitude (spatial information), mean pixel intensity (brightness), and pixel intensity standard deviation (contrast). At the single-neuron level, we correlated each neuron's response with reconstruction quality metrics (SSIM and pixel-wise correlation) across test images. We then compared these results to correlations between neuronal responses and image properties (Supplementary Figure 14). At the population level, we summarized neuronal activity for each test image using descriptors such as mean and median response amplitude, standard deviation, and skewness (Supplementary Figure 15).

Notably, despite variability among individual mice, neurons that exhibited stronger correlations with spatial information (edge magnitude) consistently showed robust negative correlations with SSIM. Conversely, neurons responsive to brightness and contrast exhibited positive and substantial correlations with SSIM and pixel-wise correlation metrics, respectively (Figure 4A). Moreover, by comparing the inter-neuron correlations with image properties against correlations among the image properties themselves, we showed that the results reflected neuron-level functional differences rather than trivial image-level statistics (Figure 4B). While brightness, contrast, and spatial information were only weakly correlated at the image level, the corresponding neuron-level correlations with reconstruction quality were substantially stronger and consistent across mice. These results suggest that V1 may process visual inputs mainly by relying on a subset of neurons specifically tuned to low-level image properties.

At the population level, the relationships between overall neural activity and reconstruction quality were more variable and less consistent than those observed at the individual level. For instance, higher mean or median responses correlated positively with better pixel-wise reconstruction metrics in mouse 21067, which, however, was not observed in the other mice. The profiles describing the distribution (skewness, mean-median ratio, etc.) were other factors related to reconstructions. However, they also displayed heterogeneous associations with reconstruction metrics across individuals (Supplementary Figure 15). These observations further emphasized that visual stimulus information may be primarily represented by distinct subsets of feature-selective neurons rather than by global population activity patterns in V1.

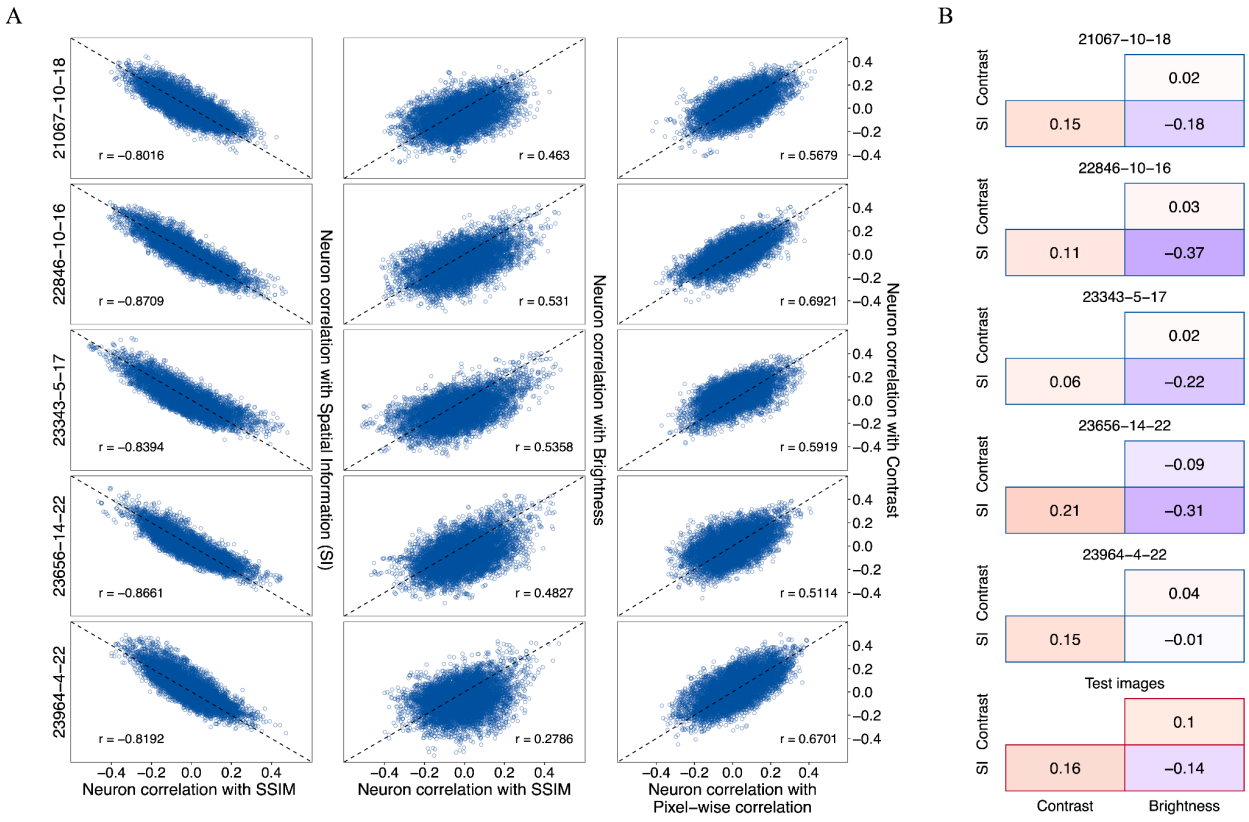

Figure 4. The relationships among reconstruction metrics, image properties, and neuron signals at the individual and population levels. Panels (A) and (B) illustrate relationships at the level of individual neurons. (A) shows the relationships between the correlation scores in neuron-to-metric (SSIM, pixel-wise correlation) and neuron-to-image property correlations (spatial information, brightness, contrast). Each point represents a single neuron. (B) The correlations among neuron sensitivities to different image properties (first five panels) and correlations among the image properties themselves (the last panel). The discrepancy between inter-neuron and inter-image property correlation values suggests that the neuron correlations reflect the neuron's functional relationships rather than the statistical dependencies among the image properties themselves.

Discussion

In this study, we introduced Sensorium-Viz, a novel neuron signal decoding framework designed to reconstruct visual inputs from the primary visual cortex of mice, specifically leveraging calcium imaging data. Our approach addresses the unique challenges of decoding responses at single-neuron levels to grayscale visual stimuli, which differ substantially from human fMRI data in previous neural decoding studies.

We first demonstrated the viability of Sensorium-Viz by evaluating it across eight quantitative metrics. Notably, our method outperforms a direct application of MinD-Vis, which was originally developed for human fMRI data. This is evidenced by higher scores in pixel-wise Pearson's correlation, SSIM, and other metrics, as well as improved visual fidelity in reconstructions. These findings additionally highlight the difficulty of directly transferring models across species and data modalities,

underscoring the importance of developing methods specifically tailored to the unique features of each experimental setup.

To place these results in context, we compared our reconstruction performance with previously reported benchmarks. In human fMRI studies, pairwise similarity based on correlation was reported as 78.1%, while SSIM values reached 62.9% and 65.3% [8,22]. By contrast, Sensorium-Viz achieved pairwise similarity scores of 91.9% in correlation and 73.1% in SSIM. These values provide a useful reference point; however, direct comparisons across modalities should be interpreted with caution. For mouse calcium imaging, Li et al. (2023) reported an SSIM of 0.5945; however, their evaluation was limited to 32×32 images. Yoshida and Ohki (2020) similarly demonstrated the feasibility of reconstructing natural images from V1 activity, reporting a median correlation of 0.43–0.45 across their image sets when similarly downsampled to 32x32 [15]. Our method achieved a comparable correlation of 0.46 on 100 unique 256x256 test images drawn from ImageNet, which spanned diverse natural scenes and minimized experimental biases. In addition, Sensorium-Viz produced reconstructions with sharper visual details compared to the blurrier outputs typically seen in earlier approaches. A direct architectural comparison was not possible, as their code and data were not publicly available despite repeated requests. Nevertheless, these results indicate that Sensorium-Viz robustly captures fine-grained visual features in calcium imaging data and extends prior work by introducing diffusion-based generative modeling.

Next, we explored the model's ability to generalize across different mice using synthetic data. By fine-tuning with only synthetic responses, Sensorium-Viz enabled a base model trained on one "source" mouse to reconstruct stimuli for a different "target" mouse. Compared to Yoshida's work [15], which reported a median correlation of 0.33 in cross-plane experiments (i.e., different scanning areas in different mice), our method achieved a median correlation of 0.40. This number indicates that incorporating synthetic responses can support more efficient cross-subject adaptation, potentially reducing the need for extensive new data collection.

One current limitation of this method, however, is that generating synthetic responses for a new mouse still requires an encoding model trained with some real data from that subject. This step is necessary to account for differences in the number of recorded neurons and scanned cortical regions across individuals[25]. From our scaling experiments with synthetic data, we estimate that training a decoder model fully on real data would require approximately 20,000–40,000 image–neuron signal pairs to achieve stable reconstruction performance (Figure 2). In contrast, training an encoder model to generate synthetic responses can be accomplished with 5,000 or fewer real images. Although we have not yet systematically determined the minimal dataset size required, our findings suggest that synthetic data can substantially reduce—though not entirely eliminate—the need for additional recordings when adapting reconstruction models to new subjects.

To demonstrate the capabilities of Sensorium-Viz as a tool for vision-decoding studies, we further examined the relationships between reconstruction qualities, image properties, individual neuron signals, and population properties across neurons. We demonstrated that individual neurons tuned to low-level image features played a crucial role in determining reconstruction quality. In contrast, the profiles describing population-level neuron activities were less predictive and more variable across mice, highlighting that V1 understands visual information more with distinct subsets of feature-selective neurons than with global population patterns. Importantly, by comparing neuron-level correlations with

correlations among the image properties themselves, we showed that these effects cannot be explained simply by trivial image statistics (e.g., brighter images being inherently easier to reconstruct). Instead, they reflect functional heterogeneity in how subsets of neurons contribute to visual decoding. Previous studies have reported similar findings, showing that a fraction of neurons make substantial contributions to visual representation in V1, while others with low explainable variance are less predictive [15,19,25]. Together, these results also suggest that identifying and selectively embedding only those neurons that make substantial contributions could also enhance the reconstruction efficiency in a diffusion-based decoder model in future work.

Another area for future improvement is reconstructing visual inputs from single-trial or individual-repeat data, which remains challenging due to the noisy and variable nature of calcium signals (Figure 1E). Similar to our approach, the previous method that achieved high decoding accuracy also relied on the availability of repeated trials[26]. Efficiently reconstructing unseen stimuli that are shown only once still requires further exploration [10].

In conclusion, our findings demonstrate that Sensorium-Viz is a powerful and flexible framework for decoding visual information from mouse cortical signals. By integrating a transformer-based diffusion model, grid-based embedding of neuronal activity, and synthetic data augmentation, we have advanced the field's capacity to reconstruct the visual stimuli from a non-primate visual system. This framework presents a novel *in silico* approach to understanding the relationship between neuronal signals in the primary visual cortex and the external visual environment.

Methods

Collect image and neuronal response datasets

Datasets were retrieved from the public repository: https://gin.g-node.org/cajal/Sensorium2022. The images and neuronal response data from 5 mice (21067, 22846, 23343, 23656, 23964) were used to develop and evaluate the image reconstruction model. The image data consisted of grayscale images from ImageNet with a shape of 144×256 (144 pixels in height and 256 pixels in width). Each image would be presented to the mice for 0.5 seconds, followed by a black screen period between 0.3 and 0.5 seconds. Neuronal response data were collected from the two-photon calcium images responding to these stimuli during the presentation periods. The signal for each neuron to each image was recorded in the accumulated relative fluorescence changes, represented by a single value. The anatomical coordinates of the neurons were also provided, representing the neuron positions relative to the pial surface. An initial partition of the train, validation, and test data was included for each mouse. The test data consisted of 100 images repeated 10 times, which were randomly mixed with the other images during the neuronal signal collection experiments.

Collect images and generate synthetic neuronal responses

Images for generating the synthetic responses were collected from the training part of the COCO 2014 Dataset (https://www.kaggle.com/datasets/jeffaudi/coco-2014-dataset-for-yolov3). The 82,784 raw images were converted to grayscale and resized to a shape of 144 × 256 to fit the input requirements of our neuronal response prediction models. We also calculated the mean and standard deviation values from

the pixels of all processed images. To generate the synthetic responses, we use the winning solution that we developed for the 2022 Sensorium Challenge, available at https://github.com/GuanLab/Sensorium2022_Challenge. A set of inputs included an image, the image after centering, a bounding box of the object in the image, and the neuron positions of the mouse. This pipeline would generate 10 response predictions for each image of each mouse. The final synthetic response was obtained by averaging the 10 predicted responses, forming an ensemble output.

Process the input data for model training and inference

The input data for training consisted of neuron positions and responses, along with the corresponding visual stimuli. The 1-channel grayscale images were first converted to 3-channel without altering colors. Each image was then resized by upscaling both sides by a factor of 256/144, followed by a center crop to produce 256x256 inputs. All channels were normalized with a mean and standard deviation of 0.5. No image augmentation techniques were applied, as neuron responses are sensitive to the content of the stimuli.

Neuron responses from real experiments were normalized using the same processes as in our previous work and the SENSORIUM challenges [19,25]. We first computed the standard deviation of individual neuron responses across the entire training set, resulting in a standard deviation vector with the same length as the responses. We then element-wise divided the raw responses by the standard deviations. Synthetic responses required no further processing, as they were generated by the model specifically designed to produce normalized responses.

Using functions from the PyInterp library, the normalized responses were further embedded into 32x32 grids, guided by neuron positions normalized to the range (-1, 1). Grid values were determined through the Inverse Distance Weighting (IDW) algorithm, considering 10 nearest neighbors. The resulting output matrices consisted of three channels: one for the interpolated response values, and two for the XY grid coordinates, which were evenly spaced between -1 and 1. The operations followed the PyInterp tutorials (https://cnes.github.io/pangeo-pyinterp/auto_examples/ex_unstructured.html).

Model architecture

Our model shares a similar structure to a latent diffusion model, consisting of the following components: (1) an AutoEncoder that encodes the input images to lower-dimensional latent space and decodes the reconstructed latent representations to final output images; (2) a denoising diffusion probabilistic model (DDPM) that iteratively diffuses and denoises input images according to the guidance from the embedded responses; and (3) an encoder that encodes the information from responses to denoise the images.

The AutoEncoder. The image data is encoded and decoded using a variational autoencoder (VAE) model with a Kullback–Leibler divergence (KL) loss, which is implemented by the AutoencoderKL module in the diffusers library. The module was loaded with the pre-trained weights of "stability/sd-vae-ft-ema" (https://huggingface.co/stabilityai/sd-vae-ft-ema), which were frozen during training.

The response projector network. The architecture of this network is inspired by the input and first block layers of Attention-Unet in the latent diffusion denoising module (Figure 1C, Supplementary Figure 2A).

The network takes in the 3-channel spatially embedded responses (shape: (3, 32, 32)), which are first encoded by a convolution layer, producing a feature map with 32 feature channels (shape: (32, 32, 32)). The resulting maps were then encoded into 64 channels through a sequential operation consisting of a group normalization with 8 groups, a SiLU nonlinearity, and a convolution layer with 64 filters. In parallel, another 64-filter convolution layer encoded the 32-channel input, and its output was added as a residual connection to the main branch (shape: (64, 32, 32)).

Next, a multi-head self-attention layer was used to encode long-range dependencies across the spatial grid. The 64-channel inputs were flattened (shape: (64, 32 × 32)), group-normalized (8 groups), and passed through an attention layer with 16 heads. The outputs were then reshaped back to the original heights and widths (shape: (64, 32, 32)) and added back to the original 64-channel inputs, forming another residual block.

Finally, the extracted feature maps were group-normalized (8 groups), SiLU-activated, and processed by a pointwise convolution to integrate the channel information in each grid (shape: (1, 32, 32)). The processed maps were then flattened into vectors and passed through a linear layer, producing conditioning vectors with a length of 1152 that represented the entire neuronal response pattern. The dimensionality of 1152 matches the predefined hidden size of the DiT-XL/2 model.

The Denoising Diffusion Probabilistic Model (DDPM). Briefly, a DDPM consists of two steps: a forward diffusion process and a learnable denoising process. During the forward diffusion process, small Gaussian noise is gradually added to the real data over T timesteps. Mathematically, given a real data point x_0, and with the reparameterization trick, the noisy data x_t at the timestep t can be expressed as:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon_t,$$
$$\text{where } \overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \text{ and } \epsilon_t \sim N(0, I) \quad (1)$$

In implementation, α_s is calculated from $1 - \beta_s$, where β_s is retrieved from a linear space with 1000 data points from 0.0001 to 0.02 (a linear scheduler with 1000 diffusion timesteps). As $T \to \infty$, x_T becomes equivalent to a point sampled from an isotropic Gaussian distribution.

During the reverse denoising process, a denoising function is trained to learn a parameterized distribution that approximates the reverse of the forward process, enabling the recovery of clean samples from noisy inputs. Given x_t, this process estimates the conditional distribution of x_{t-1} at the previous timestep:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

where μ_θ and Σ_θ are statistics that can be learned by a neural network, which, in our method, is the diffusion transformer (DiT). In practice, one way to estimate p_θ in DDPM is training a neural network $\epsilon_\theta(x_t, t)$ to approximate the real noise ϵ_t added at the timestep t instead of μ_θ. Once p_θ trained, new images can be sampled by initializing $x_{tmax} \sim N(0, I)$ and repeatedly sampling $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$.

The Diffusion Transformer (DiT). DiT shares a similar network structure as the ViT, including a patch embedder (patch size = 2) and a sin-cos position embedder to patchify the inputs—the latent images encoded by the autoencoder—into image tokens. These tokens are then processed by 28 DiT blocks, which incorporate conditioning information from the neuron responses and timesteps. A final layer decodes the tokens and unpatchifies them to produce two outputs: a noise prediction and a diagonal covariance prediction. Both outputs have the same shape as the input.

Within the DiT blocks, the timestep is encoded by a sinusoidal timestep embedder and added to the encoded neuron response information, forming the conditions that guide the current denoising step. Conditions were incorporated into the blocks through an adaptive layer norm (adaLN)-Zero module. Inside this module, a linear encoder maps the conditions to six tensors: $(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)$. Each has the same shape as the input conditions. The six tensors are applied to a transformer block through scale and shift operations. Mathematically, given an input x, a multi-head self-attention layer MSA, and a point-wise feed-forward layer MLP, the operation in one block can be written as the following, ignoring the normalization steps:

$$x_{msa} = MSA(x * (1 + \gamma_1) + \beta_1) * \alpha_1 + x \quad (3)$$

$$x_{out} = MLP(x_{msa} * (1 + \gamma_2) + \beta_2) * \alpha_2 + x_{msa} \quad (4)$$

All the multiplications in these formulas are element-wise. A similar conditioning step was also applied in the final layer, where γ and β were generated to scale and shift the inputs.

where γ and β were generated to scale and shift the inputs.

The training losses

The model is trained to optimize both the predicted noises and the learned reverse process covariance. The loss between the real Gaussian noise and the prediction at the timestep t is defined by a mean squared error (MSE):

$$L_{simple} = ||\epsilon_\theta(x_t, t) - \epsilon_t||_2^2 \quad (5)$$

For the covariance, the model minimizes the following variational lower bound loss given by the KL-divergence D_{KL} between the true posterior distribution defined by the forward process q and the learned reverse process p_θ at $t \geq 1$:

$$L_{vlb} = D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \quad (6)$$

Since both q and p_θ are Gaussian, the divergence can be calculated from the mean and variance of these two distributions ($\mu(t)$, $\Sigma(t)$, $\mu_\theta(t)$, $\Sigma_\theta(t)$). Using the α, β defined in the previous DDPM section (eq. 1), we can derive the posterior mean μ and log variance Σ by:

$$\mu(t) = \frac{\beta_t\sqrt{1-\overline{\alpha_{t-1}}}}{1-\overline{\alpha_t}}x_0 + \frac{(1-\overline{\alpha_{t-1}})\sqrt{\alpha_t}}{1-\overline{\alpha_t}}x_t \quad (7)$$

$$\Sigma(t) = log(\frac{\beta_t(1-\overline{\alpha_{t-1}})}{1-\overline{\alpha_t}}) \quad (8)$$

The mean $\mu_\theta(t)$ of p_θ can be calculated from the learned noise $\epsilon_\theta(t)$:

$$\widehat{x_0} = \sqrt{\frac{1}{\overline{\alpha_t}}}x_t - \sqrt{\frac{1}{\overline{\alpha_t}} - 1}\epsilon_\theta(t) \quad (9)$$

$$\mu_\theta(t) = \frac{\beta_t\sqrt{1-\overline{\alpha_{t-1}}}}{1-\overline{\alpha_t}}\widehat{x_0} + \frac{(1-\overline{\alpha_{t-1}})\sqrt{\alpha_t}}{1-\overline{\alpha_t}}x_t \quad (10)$$

while the log variance $\Sigma_\theta(t)$ is computed from the covariance output $\Sigma_\theta^*(t)$ of the final DiT layer:

$$\Sigma_\theta(t) = \frac{\Sigma_\theta^*(t)+1}{2}log(\beta_t) + (1 - \frac{\Sigma_\theta^*(t)+1}{2})\Sigma(t) \quad (11)$$

When $t = 0$, L_{vlb} reduces to $-p(x_0|x_1)$, which corresponds to the log-likelihood of a Gaussian distribution, which is parameterized by $\mu_\theta(1)$, and $0.5 * \Sigma_\theta(1)$, discretizing to x_0. The total loss used to train the model is given by $L = L_{simple} + L_{vlb}$.

The classifier-free guidance (CFG)

Conditional diffusion models take extra input, such as embedded text, labels, or, in our case, neuron response, as guidance for image generation. To encourage the model to better understand and rely on these conditions, previous works employ a technique known as classifier-free guidance, in which some samples are intentionally created without including the condition during training. This method teaches the model to distinguish between conditioned and unconditioned outputs, allowing it to use the guidance more effectively when it is provided during generation. This technique is widely used and generally yields significantly better sample qualities. In the image-class-guided generation, a CFG sample has an additional "null" class other than the actual classes.

To utilize the CFG technique in our neuron signal-guided reconstruction, we created "null" responses by setting the first embedded response channel (the channel for neuron responses) to 0 while keeping the

position information in the other two channels. The dropout rate was 0.1, meaning that 10% of the training data in each batch were assigned "null" responses.

Devices and settings for training the models

The models were trained on one or multiple NVIDIA L40S 48G GPUs, with a batch size of 32 per GPU and 80,000 training steps. The parallel training across multiple GPUs was handled by the "DistributedDataParallel" (DDP) module of PyTorch. Losses were optimized by the "AdamW" optimizer with a learning rate of 5e-5, default Adam betas (0.9, 0.999), and 0 weight decay. To stabilize the training process, a gradient clipping step was introduced during weight updates, which limits the gradient norm of the parameters to a maximum of 1.0 (max_norm = 1.0).

Our preliminary experiments observed abnormal loss changes, where the losses could suddenly increase and explode to NaN, which may be due to outliers in both the real and synthetic data. To eliminate these samples, we trained a model with all samples from the five mice and monitored the loss changes by the exponential moving average (EMA) of the losses. Batches with suspected outliers would be saved when the loss exceeded 3 times the EMA or became NaN. A total of 1809 outliers (1694 synthetic and 115 real) were excluded from the training set.

Reconstruction with the neuron responses

After training a model, we generate the reconstructions from random noises in the latent space and under the guidance of the embedded neuron responses. The noises had the shape of (4, 32, 32), where "4" is the channel number and the latter two values represent the width and height. These numbers are the shapes of the auto-encoder outputs.

We also applied the CFGs in the inference stage. The initial latent image would be repeated twice to generate reconstructions under real and null conditions, and the input conditions also consisted of real and null neuron responses, as prepared above. The output of the model (ϵ_θ) which estimates the real noise added at the timestep t in the diffusion steps, would be derived from the unconditioned ($\epsilon_{\theta,\, uncond}$) and conditioned ($\epsilon_{\theta,\, cond}$) estimations with a CFG scale (4.0 in our implementation):

$$\epsilon_\theta = \epsilon_{\theta,\, uncond} + CFG * (\epsilon_{\theta,\, cond} - \epsilon_{\theta,\, uncond}) \qquad (12)$$

Then, with the mean and log variance we derived from formulas (10) and (11), we gradually generated the reconstruction by:

$$x_{t-1} = \mu_\theta(t) + exp(0.5 * \Sigma_\theta(t)) * \epsilon \quad \epsilon \sim N(0, I) \qquad (13)$$

The noise term will be removed when $t = 0$.

Evaluation metrics

The performance of reconstructions is evaluated using both pixel-level and overall image content metrics, including mean squared errors (MSE), Pearson's pixel-wise correlation, learned perceptual image patch similarity (LPIPS), and structural similarity index (SSIM). All the metrics receive a pair of images (the ground truth I_{gt} and the reconstruction I_{recon}) as inputs, where the pixel values range from 0 to 1. A matrix of Gaussian noise with the same shape as I_{recon} is used to calculate the random baseline of each metric. The baseline values for the metrics are around 0.14 (MSE), 0 (pixel-wise correlation), 0.846 (LPIPS), and 0.007 (SSIM), respectively. The baseline values for the pairwise similarities are all 0.5.

MSE and pixel-wise correlation are measured pixel by pixel and implemented using their formulas and the functions from the Python NumPy package. SSIM is implemented with the "structural_similarity" function from scikit-image [27]. The "data_range" parameter in this function is set to the difference between the maximum and minimum pixel values I_{recon}. LPIPS computes the similarity between the activations of images from a predefined network, which is implemented using "LearnedPerceptualImagePatchSimilarity" from TorchMetrics with a pre-trained VGG-16 network (net_type = "vgg") [28]. Additionally, we employ pairwise similarity comparisons based on these metrics to assess accuracy, as used in previous research. For an inference, it will be compared with the other gold standards, and the metrics will be calculated (Figure 2B). The pairwise similarity measures how specific the inferences are reconstructed for the gold standard.

Application of MinD-Vis to mouse neuron responses

To compare our model with current state-of-the-art techniques for reconstructing visual stimuli from brain signals, we applied the techniques used in MinD-Vis to neuron responses [10].

MinD-Vis is a latent diffusion model trained in two separate stages. The first stage, referred to as "sparse-coded masked brain modeling" (SC-MBM), involves training a masked autoencoder to reconstruct masked signals and embedding the input neuron signals in a latent space. Our input neuron responses, both real and synthetic, were ordered by their anatomical positions as provided in the raw data. The mask ratio was 0.75. To ensure non-negative reconstructions, we wrapped the original output in an ELU activation with a +1 offset. All the other settings were kept the same as MinD-Vis suggested. We trained for 400 epochs using the standard MSE loss between the gold standard and reconstructed responses on the training data. We then fine-tuned the embeddings on the test responses for an additional 25 epochs.

The second stage, referred to as "double-conditioned latent diffusion modeling" (DC-LDM), utilizes the learned encoder from the first stage as conditioning to fine-tune a latent diffusion model. We also followed the same settings as MinD-Vis and trained on our datasets for 280 epochs for each mouse. We saved the model every 20 epochs. The weights that achieved the best pixel-wise correlation were loaded for benchmarking and making inferences on each mouse.

Code availability

The codes of this work are available at: https://github.com/GuanLab/sensorium-viz.

Data availability

Datasets of the mice can be retrieved from https://gin.g-node.org/cajal/Sensorium2022. Synthetic responses for the COCO images and the model weights needed to reproduce the paper results can be obtained from Google Drive:
https://drive.google.com/drive/folders/1GbJ7V2AzVezKW3U0lwhrKYaeQni2ntef

References

1. Teng, C. & Kravitz, D. J. Visual working memory directly alters perception. *J. Vis.* **19**, 81c (2019).
2. Thirion, B. *et al.* Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage* **33**, 1104–1116 (2006).
3. Miyawaki, Y. *et al.* Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* **60**, 915–929 (2008).
4. Fujiwara, Y., Miyawaki, Y. & Kamitani, Y. Modular encoding and decoding models derived from bayesian canonical correlation analysis. *Neural Comput.* **25**, 979–1005 (2013).
5. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* **8**, 15037 (2017).
6. Roelfsema, P. R., Denys, D. & Klink, P. C. Mind reading and writing: The future of neurotechnology. *Trends Cogn. Sci.* **22**, 598–610 (2018).
7. Ozcelik, F., Choksi, B., Mozafari, M., Reddy, L. & VanRullen, R. Reconstruction of perceived images from fMRI patterns and semantic brain exploration using instance-Conditioned GANs. *arXiv [cs.CV]* (2022) doi:10.48550/ARXIV.2202.12692.
8. Fang, T., Qi, Y. & Pan, G. Reconstructing perceptive images from brain activity by shape-semantic GAN. *arXiv [cs.NE]* (2021).
9. Lin, S., Sprague, T. & Singh, A. K. Mind Reader: Reconstructing complex images from brain activities. *arXiv [q-bio.NC]* (2022) doi:10.48550/ARXIV.2210.01769.
10. Chen, Z., Qing, J., Xiang, T., Yue, W. L. & Zhou, J. H. Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding. *arXiv [cs.CV]* (2022).
11. Takagi, Y. & Nishimoto, S. High-resolution image reconstruction with latent diffusion models from human brain activity. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2023). doi:10.1109/cvpr52729.2023.01389.
12. Diamond, M. E., von Heimendahl, M., Knutsen, P. M., Kleinfeld, D. & Ahissar, E. ‘Where’

and ‘what’ in the whisker sensorimotor system. *Nat. Rev. Neurosci.* **9**, 601–612 (2008).

13. Sofroniew, N. J., Cohen, J. D., Lee, A. K. & Svoboda, K. Natural whisker-guided behavior by head-fixed mice in tactile virtual reality. *J. Neurosci.* **34**, 9537–9550 (2014).
14. Franke, K. *et al.* State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature* **610**, 128–134 (2022).
15. Yoshida, T. & Ohki, K. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nat. Commun.* **11**, 872 (2020).
16. Malathi, T. & Bhuyan, M. K. Performance analysis of Gabor wavelet for extracting most informative and efficient features. *Multimed Tools Appl* **76**, 8449–8469 (2017).
17. Garasto, S., Bharath, A. A. & Schultz, S. R. Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex. *bioRxiv* (2018) doi:10.1101/300392.
18. Li, W. *et al.* The brain-inspired decoder for natural visual image reconstruction. *Front. Neurosci.* **17**, 1130606 (2023).
19. Willeke, K. F. *et al.* The Sensorium competition on predicting large-scale mouse primary visual cortex activity. *arXiv [q-bio.NC]* (2022).
20. Peebles, W. & Xie, S. Scalable diffusion models with transformers. *arXiv [cs.CV]* (2022).
21. Li, B. M., Cornacchia, I. M., Rochefort, N. L. & Onken, A. V1T: large-scale mouse V1 response prediction using a Vision Transformer. *arXiv [cs.CV]* (2023).
22. Shen, G., Dwivedi, K., Majima, K., Horikawa, T. & Kamitani, Y. End-to-end deep image reconstruction from human brain activity. *Front. Comput. Neurosci.* **13**, 21 (2019).
23. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018). doi:10.1109/cvpr.2018.00068.
24. Betker, J. *et al.* Improving image generation with better captions.
25. Kaiwen Deng, Peter S. Schwendeman, Yuanfang Guan. Predicting Single Neuron

Responses of the Primary Visual Cortex with Deep Learning Model. *Adv. Sci.* **2305626**, (2024).

26. Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).
27. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).
28. Detlefsen, N. *et al.* TorchMetrics - Measuring Reproducibility in PyTorch. *J. Open Source Softw.* **7**, 4101 (2022).